COLOR-AND TEXTURE-BASED SALIENT MAP

Stéphane Paris

COLOR- AND TEXTURE-BASED SALIENT MAP

Stéphane Paris

LITA, Paul Verlaine University, Metz, France

ABSTRACT

Salient maps are often obtained by extracting color- and texture-based characteristics ; that is, identify dominant textures and colors and associate them with regions. In this paper, dominant textures and colors are identified by means of the k-means. Then, connected components are extracted. This two step processing delivers a hierarchy of salient maps. Moreover, it uses a similarity measurement taking into account texture and color specificities.

Index Terms— descriptor, color, texture, salience map.

1. INTRODUCTION

The salient map of an image identifies the present objects [1, 2, 3, 4, 5, 6, 7]. It is a synthesis where the insignificant objects are deleted and where the silhouettes of the retained object do not need to be precise. Since the segmentation delivers every segments with precise contours, the salient maps and the segmented images are different. The embedded processes are strongly similar in the sens that the pixel characterization by texture- and color-based vector descriptors are the same. But their used is a little bit different[8, 9].

This paper presents a salient map extraction process (SME) built on with the k-mean algorithm followed by a connected component extraction. However, it results in a hierarchy of salient maps.

2. RELATED WORKS

With the growing of digital multimedia data, salient maps are strongly studied [1, 2, 10, 9].

Blobworld is one of the pioneered projects based on salient maps [1]. The descriptors are six-dimension vectors defined by color and texture. Every region is characterized by its own mean texture value, by its two most predominant color values and by some geometrical information. The segmented image is either stored into an image database or used as a request to retrieve similar images in the database. The scale-space technique used by Blobworld is tuned for every pixel. This process requires the image gradient analysis to be controlled by an aperture defining the neighborhood [11]. The moment of order-two and the polarity are measured at every pixel. The polarity of a pixel p can be understood as the ratio of the orientation averages in its neighborhood between the pixels roughly in the orientation of pixel p and the pixels in a different orientation. To avoid such calculations, SME quantifies the texture with the Haar transform. Like for all wavelet transforms, the delivered hierarchical scale-space informs on the local properties of the image. As Blobworld descriptors describe color and texture, the distance is based on two spaces without relation. To overcome to this drawback, SME uses a combination of two distances, one for the color-space and the other for the texture-space.

Simplicity is very similar to SME. It partitions descriptors with a k-means algorithm. The Haar transform delivers the texture information but in a single level decomposition [2]. However, Simplicity focuses on image categorization : photography versus graphic. To observe significant texture patterns, i.e. deep-structures [11, 12], SME uses a multilevel decomposition.

Other projects make spatial and descriptive groupings simultaneously [13, 14, 8, 9]. But, before merging the two processes, SME proposes to observe the performance of each one separately. In this way, one can be sure that the regions of salient maps are *stable*. By delivering a salient map hierarchy, SME allows to observe regions stable over a range of scales.

3. OVERVIEW

SME operates in three successive stages separating the texture description and color description from the spatial grouping : (1) pixels are characterized according to their color and texture information (sec. 4 and 5) ; (2) descriptors are partitioned (sec. 6) ; (3) Connected component are extracted (sec. 7). The first stages characterize the pixels and the last one spatially regroups pixels into salient regions.

4. COLOR DESCRIPTORS

The color space is chosen to separate luminance from chrominance. We have experienced two color models : (1) the Gaussian model [15], (2) the CIE Lab model [16]. Experiments exhibit a more accurate description with the Gaussian model than with the Lab model. Moreover, the human visual system (HSV) is well approximated by the Gaussian model, while offering a linear transform with the RBG space, unlike the Lab

model. Next a two component vector describes every pixel with the energy of the two color channels. A more detailed discussion is proposed in section 8.

5. TEXTURE DESCRIPTORS

At every pixel, color and texture are measured separately. We use repetitive pattern as texture definition. These patterns vary in size and complexity with the observed textures. Following this definition, a multi-scale approach seems well adapted [11, 12]. In this way, a pixel is related to its neighborhood, allowing common patterns to be identified[17, 18, 19, 20]. Thus, we use the Haar wavelet transform for texture identification. It delivers a treelike decomposition with a selection of the best sub-bands. The maximum level J of decomposition is a parameter – set to 3 during experiments. At start, the image is separated in one sub-band, A, of approximation and three sub-bands, H, V and D, of respectively horizontal, vertical and diagonal details. Next, every sub-band is iteratively decomposed until either the maximum level J of decomposition is attained or the decomposition criterion is unverified before. A node in the wavetree (i.e; a sub-band) is a leaf if either its level of decomposition is J or its entropy is less than the quarter of the sum of the entropies of its sub-bands [21]. So, the components of vector descriptors are the coefficients of the leaves.

6. CLUSTERING

A K-means technique groups the descriptors into K clusters. Since the convergence can lead to a local minimum, the initialization is of importance. However, rather than the exact partition, the clustering has to deliver the main clusters, i.e. those which are well separated and relatively compact. Thus, the centers can be initialized randomly. It remains three parameters to be estimated precisely : (1) What distance to use ? (2) What *cluster separation criterion* to use ? (3) What *compactness criterion* to use ? The two criteria allow to estimate the number K of clusters and they ensure that the random initialization of the cluster centers is not too harmful. The Euclidean distance between two texture descriptors

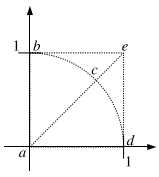


Fig. 1. Illustration of the anisotropic Euclidean distance.

$x = (x_1, \cdots, x_n)^T$ and $y = (y_1, \cdots, y_n)^T$:

$$d_E^2(x, y) = \sum_{i=1}^{n} (x_i - y_i)^2$$

does not inform about the information delivered by the Haar transform. It hypothesizes an isotropic space whereas the wavelet decomposition provides the horizontal, vertical and diagonal details. Thus, it is more pertinent to use an anisotropic Euclidean distance which takes into account the orientation information of the wavelet decomposition. For example, figure 1 shows five 2D points ($n = 2$). The Euclidean distances from point a to points b, c and d are equal and the distance to point e is greater :

$$d_E^2(a, b) = d_E^2(a, c) = d_E^2(a, d) = 1 \text{ and } d_E^2(a, e) = 2$$

In other words, if the information delivered by the two axes have to be considered, the Euclidean distance is not appropriate. But, following the study of Wang et al. [22], we can use an anisotropic Euclidean distance :

$$d_I^2(x, y) = \sum_{i,j=1}^{n} g_{i,j}(x_i - y_i)(x_j - y_j)$$

where $g_{i,j}$ is a continuous function positive and decreasing monotonically. For example : $g_{i,j} = \frac{1}{2\pi\sigma^2} e^{-\frac{(i-j)^2}{2\sigma^2}}$. With $\sigma = 1$ and neglecting the normalization constant, we obtain : $g_{i,j} \approx e^{-\frac{(i-j)^2}{2}}$. In this way, the descriptors, which share identical sub-band coefficients, are close together. For the precedent example, the distances d_I are :

$$d_I^2(a, b) = d_I^2(a, d) = 1, d_I^2(a, c) = 1 + e^{-\frac{1}{2}} \text{ and } d_I^2(a, e) = 2\left(1 + e^{-\frac{1}{2}}\right)$$

If we consider that the abscissa informs on the approximation coefficients and that the ordinate informs on the detail coefficients, the distances d_I shows that points b et d are closer to point a than is point c. In other words, the distance d_I emphasizes that the approximation coefficients (resp. details coef.) of b (resp. d) and a are identical, whereas c has no common coefficient with the point a.

For texture definition through pattern extraction, it is convenient to identify pixels which share common sub-band coefficients. So, from now, the distance d_I is used to regroup texture descriptors and the distance d_E to group together color descriptors based on energy measurements.

The similarity between two pixels u and v is then the combination of these two distances. If T_u (resp. T_v) is the texture descriptor of pixel u (resp. v) and C_u (resp. C_v) its color descriptor, the similarity can be measured as a linear combination of these two distances : $d(u, v) = \alpha\, d_I(T_u, T_v) + (1 - \alpha)\, d_E(C_u, C_v)$ where α is a constant between 0 and 1 which weights the relative importance of the texture to the color into the observed image. For example, α can be the ratio between the texture energy and the color energy :

$$\alpha = \frac{E_T}{E_C + E_T} \textbf{ with } \left\{ \begin{array}{rcl} E_T & = & \sum_{i,j=1}^{N,M} ||T_{u_{i,j}}||^2 \\ E_C & = & \sum_{i,j=1}^{N,M} ||C_{u_{i,j}}||^2 \end{array} \right.$$

where N and M are the size of the original image and $T_{u_{i,j}}$ (resp. $C_{u_{i,j}}$) is the texture (resp. color) descriptor of pixel $u_{i,j}$.

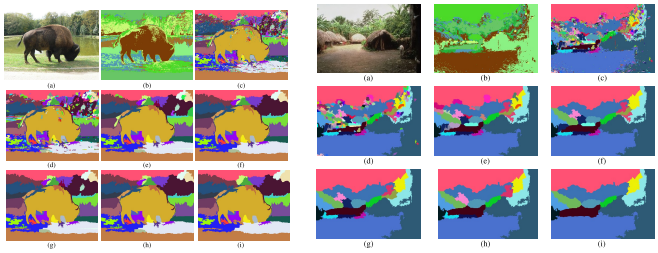
The number K of clusters is iteratively estimated. Several clustering are done. The best is the one for which the compactness and separation criteria are the highest. So, the criterion to minimize evaluates the separation $q_{i,j}$ between clusters i and j weighted by their compactness p_i and p_j :

$$M = \prod_{i,j=1,i+1}^{K-1,K} p_i p_j q_{i,j} \text{ with } \begin{cases} p_i = \frac{\sigma_i}{\sum_j \sigma_j} \\ q_{i,j} = \frac{d(\mu_i,\mu_j)}{\sum_{m,n=1,m+1}^{K-1,K} d(\mu_m,\mu_n)} \end{cases}$$

where μ_k is the mean of vectors x_k belonging to the cluster $\mathcal{C}_k$ and σ_k is its standard deviation. At start, three clustering are done with the number of clusters sets to K, $K+1$ and $K-1$ respectively. If the best is that with K clusters, the process stops. Otherwise, the search for the best continues by increasing or decreasing the number of clusters according to the number of clusters ($K+1$ or $K-1$) of the clustering which minimizes the criterion. To ensure the end of the process, the number has to belong to the range : $2 \leq K \leq 20$.

7. SALIENT REGION EXTRACTION

Once the descriptors are clustered, the salient regions can be identified. They are defined by pixels spatially neighbors with descriptors belonging to the same cluster. Thus, a classical 8-nieghbors connected component extraction is used. Next, the connected components estimated too small are suppressed. To do that, the connected components histogram H which informs about the size of the connected components is cut off. Experiments show that H has two kinds of connected components quite well separated ; small connected components can be discarded while few significant connected components can be dilated. An automatic thresholding is used : $T_H = \mu_H + \gamma\,\sigma_H$ with μ_H and σ_H the mean and the standard deviation of H and γ a weighting parameter.



(I) **(II)**

Fig. 2. **(a)** The original image ; **(b)** the clustering map ; **(c)** the cc map without cleaning ; **(d-i)** the cc map with $\gamma = 0.0, 0.2, 0.4, 0.6, 0.8$ and 1.0.

8. EXPERIMENTATION

Database images come from Internet and personal albums. Every experiment results in a clustering map and in several connected component maps with different values of γ. [1] Parameters α and γ are specifically discussed.

8.1. parameter α

Roughly speaking, the experiments show the effectiveness of parameter α : i.e., it varies according to the relative importance of the texture information. The salient maps of figures 2.I and 2.II exhibit the relevant objects of the scenes. These objects are quite well identified and not so badly localized.

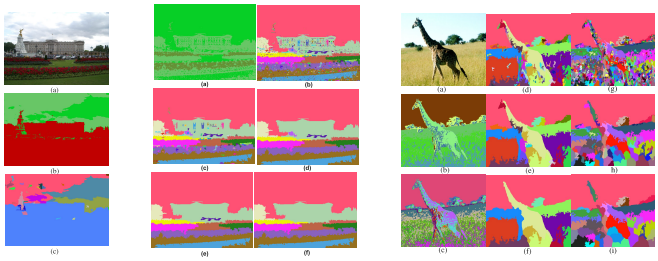
But, sometimes, these texture- and color-based energies ratios do not inform on the reality. And so, the clustering is affected. For example, the Buckingham palace picture is constituted of several depths (see Fig. 3.I.a). The foregrounds of red flowerbeds are obviously dominated by color information when the backgrounds of buildings framed by trees are rather texture-based informative. But the parameter α is set to 0.91861 for the whole image saying that the texture is dominant even for the foregrounds[2]. So we need to locally adapt the α measurement. We change the global α measurement by a set of local, pixel-based, measurements; i.e., for every pixel x and its spatial neighborhood $\mathcal{V}_x$ the following rule is used :

$$\alpha_x = \frac{E_{T,x}}{E_{T,x} + E_{C,x}} \text{ with } \begin{cases} E_{T,x} &= \frac{1}{|\mathcal{V}_x|} \sum_{y \in \mathcal{V}_x} y_T^2 \\ E_{C,x} &= \frac{1}{|\mathcal{V}_x|} \sum_{y \in \mathcal{V}_x} y_C^2 \end{cases}$$

It follows a modification of the distance calculation. The distance of a vector x to a cluster $\mathcal{C}$, with texture center μ_T and color center μ_C, is now defined by :

$$d(x,\mathcal{C}) = \alpha_{\mathcal{C}}\, d_T(x,\mu_T) + (1-\alpha_{\mathcal{C}})\, d_C(x,\mu_C)$$

where $\alpha_{\mathcal{C}}$ is the mean value of the α_y values of all vectors y belonging to cluster $\mathcal{C}$. Using this new definition, figure 3.II shows the salient maps of the picture of Buckingham palace : Fig 3.II.a shows the cluter map ; Fig. 3.II.b its cc map without cleaning ; Fig. 3.II.c (resp. d and e) is the cc map obtained for $\gamma = 0.0$ (resp. 0.2 and 0.4) ; Fig. 3.II.f is obtained for $\gamma \in [0.6, 1]$.



(I) global α **(II)** local α **(III)** global vs local α

Fig. 3. See text.

8.2. parameter γ

During experiments, parameter γ is set to 0.0, 0.2, 0.4, 0.6, 0.8 and 1.0. This delivers a hierarchy of salient maps easy

[1] For a colored presentation, see **www.lita.univ-metz.fr/∼paris/Conf10/**.

[2] see cluster map (Fig 3.I.b) and cc map (Fig. 3.I.c).

to calculate since a salient map can be obtained from its preceding one – with a lower value of γ. This can be data for a process like LabelME [23]. However, the giraffe example (Fig. 3.III) shows that, observing the foreground, color information is local while texture information is global and that the background is textured and well segmented. In fact, the fixed level of decomposition $(J = 3)$ does not allow the perception of the giraffe texture. The comparition between the global (Fig. 3.III.b) and the local (Fig.3.III.c) estimates of α shows much details are observed with the local estimate ; but the head disapears in both case : Fig. 3.III.d, f and e are the cc maps from global estimate and Fig. 3.III.g, h and i from local estimate. Moreover, the separation between luminance and chrominance is not complete; shadows yet influence the color information. In spite of the efficiency of the Gaussian color space relative to the Lab space, we have to use a more effective (often more complex) color space [24, 25].

9. CONCLUSION

With the help of the new distance, the clustering process is more effective. During the experiments, the k-means algorithm estimates the clusters and their number. Although it is a slow and heavy process, it involves no learning stage and only few and generic hypotheses. But, a more efficient algorithm can be investigated : e.g., mean-shift [8], MSER[26]. Moreover, the connected component extraction embeds a cleaning step to remove too small connected components. So, it delivers a hierarchy of salient maps. In spite of the fact that region-based CBIR has been proved to be sensitive to clutters and occlusions, a hierarchy of salient maps can be part of a hierarchical visual vocabulary [27]. According to the author, no study proposes a framework to estimate the relative flexible localization of *affine regions* and/or *blobs*. May be regions are part of this framework with the CBIR goal directed by the object to retieve [28].

10. REFERENCES

[1] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld : Image segmentation using expectation-maximization and application to image querying," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1026–1038, August 2002.

[2] J. Z. Wang, J. Li, and G. Wiederhold, "Simplicity : Semantic-sensitive integrated matching for picture libraries," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 23, no. 9, pp. 947–963, September 2001.

[3] Y. Chen and J. Z. Wang, "Imaghe categorization by learning and reasonnig with regions," *Journal of Machine Learning Research*, vol. 5, pp. 913–939, August 2004.

[4] P.-J. Cheng and L.-F. Chien, "Effective image annotation for searches using multilevel semantics," *International Journal of Digital Libraries*, vol. 4, pp. 258–271, 2004.

[5] P.W. Huang and S.K. Dai, "Design of two-stage content-based image retrieval system using texture similarity," *Information Processing and Management*, vol. 40, pp. 81–96, 2004.

[6] B.S. Manjunath and W.Y. Ma, "Texture features for browsing and retieval of image data," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, August 1996.

[7] P. L. Rosin, "Edges : Saliency measures and automatic thresholding," *Machine Vision Application*, vol. 9, no. 4, pp. 139–159, 1997.

[8] D. Comaniciu and P. Meyer, "Mean shift : A robust approach toward feature space analysis," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[9] Paul L. Rosin, "A simple method for detecting salient regions," *Pattern Recognition*, vol. 42, pp. 2363–2371, 2009.

[10] D. Nistér and H. Stewénius, "Linear time maximally stable extremal regions," in *European Conference on Computer Vision*, 2008, vol. 2, pp. 183–196.

[11] J.J. Koenderink and A.J. van Doorn, "Image structure," in *DAGM-Symposium*, 1997, pp. 3–35.

[12] T. Lindeberg, *Scale-space theory in computer vision*, Kluwer academic publisher, 1994.

[13] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 2000, no. 22, pp. 888–905, August 8.

[14] S. Wang and J. Siskind, "Image segmentation with ratio cut," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 25, no. 6, pp. 675–690, June 2003.

[15] M. A. Hoang, J. M. Geusebroek, and A. W. M. Smeulders, "Color texture measurement and segmentation," *Signal Processing*, vol. 85, no. 2, 2005.

[16] R. S. Berns, *Billmeyer and Saltzman's principles of color technology*, 3ième édition, Wiley, 2000.

[17] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[18] W.-Y. Ma and B. S. Manjunath, "Netra : a toolbox for navigating lage image databases," *Journal of Multimedia Systems*, vol. 7, pp. 184–198, 1999.

[19] Mikolajczyk K., Tuytelaars T., Schmid C., A. Zisserman, Mattas J., Schaffalitzky F., Kadir T., and Van Gool L., "Comparision of affine regions detectors," *International Journal of Computer Vision*, vol. 65, no. 1, pp. 43–72, 2005.

[20] J. van de Weijer and C. Schmid, "Coloring local feature extraction," in *European Conference on Computer Vision*, 2006, pp. 334–348.

[21] R.R. Coifman and M.V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Transaction on Information Theory*, vol. 2, no. 38, pp. 713–718, 1992.

[22] Wang L., Zhang Y., and Feng J., "On the euclidean distance of images," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1334–1339, August 2005.

[23] J. C. Russell and A. Torralba, "Labelme : a database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, May 2008.

[24] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluation of color descriptors for object and scene recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. To appear, 2009.

[25] T. Gevers, J. van de Weijer, and H. Stokman, *Color Feature Detection*, chapter 10, pp. 203–226, L. Rastislav and K. N. Plataniotis, 2006.

[26] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *British Machine Vision Conference*, 2002, pp. 384–393.

[27] Nistér D. and Stewénius, "Scalable recognition with a vocabulary tree," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 2161–2168.

[28] A. Y.-S. Chia, S. Rahardja, D. Rajan, and M. K. H. Leung, "Structural descriptors for category level object detection," *IEEE Transaction on Multimedia*, vol. 11, no. 8, pp. 1407–1421, December 2009.